'Lexical teddy bears' and the processability of learner language by native speakers

MILLAR Neil

‘Lexical teddy bears’ and the processability of learner language by native speakers

ABSTRACT

Second language learners tend to rely ‘Lexical teddy bears’ – high frequency, high utility words that combine readily (e.g. *good, big, bad, nice*). Although this strategy allows a speaker to express a wide range of meanings, such ‘teddy bears’ lack the predictive power of a more precise lexical choice and result in weak collocations. Applying the theory of Lexical Priming, the present study investigates how this feature of learner language impacts on online processing by native speakers. A self-paced reading experiment was used to compare the reading times for weakly collocating adjective-noun bigrams and semantically equivalent strongly collocating bigrams. The results indicate that nouns preceded by weakly collocating adjectives (e.g. *good memories*) require more processing effort than the same nouns when preceded by strongly collocating adjectives (e.g. *fond memories*). This finding supports claims that collocation offers processing advantages. Implications for the role of collocation in language learning are discussed.

INTRODUCTION

It is widely recommended that second-language instruction should ensure that learners develop a rich repertoire of formulaic sequences, including collocation - combinations of words that habitually co-occur. If this is justified, it follows that collocation use should have some impact on communication. However, it seems that few researchers have sought to objectively evaluate how learner use of collocation impacts on online processing.

Mastery of collocation norms is especially difficult (Wray 2000). One strategy that learners often employ is the use of so-called Lexical Teddy Bears - high frequency, high utility polysemous words (e.g. *good, big, bad, nice* etc.) that combine readily where a more precise strongly collocating word could also be selected - for example, *good memories* vs. *fond memories*, *a big argument* vs. *a heated argument*, or *a good career* vs. *a distinguished career*. According to Lexical Priming theory (Hoey 2005), and other usage-based theories, this strategy has implications for processing and, therefore, albiet indirectly, communication. A reading study was, therefore, undertaken to assess how learner reliance on Lexical Teddy Bears impacts on online processing by native speakers.

The structure of the paper is as follows. The following section provides a brief overview of Lexical Priming theory (Hoey 2005) and learner use of collocation. The subsequent sections outline the hypothesis to be tested, the methodology and the results. The results are discussed with regards to Lexical Priming theory and some tentative implications for our understanding of second language communication are proposed.

BACKGROUND

It has long been recognised that the way in which we use language is not as creative as once thought. As Pawley and Syder put it "native speakers do *not* exercise the creative potential of syntactic rules to anything like their full extent" (1983:193, emphasis original). It seems rather that we prefer to re-use fixed and semi-fixed chunks of language with which we (i.e. the speaker and the hearer) are already familiar. This is perhaps best exemplified by collocation as observed in text corpora. Corpus linguistics has demonstrated how words are consistent in their co-occurrence preferences - not just co-occurrence with particular lexical items (i.e. lexical collocation), but also with certain meanings (semantic preference and prosody), grammatical functions (colligation) and discourse roles (Sinclair 1992, Hoey 2005).

Hoey (2005) proposes that such co-occurrence patterns observed in corpora are the textual realisation of a "psycholinguistic phenomenon" which he refers to as *Lexical Priming*. His notion of 'priming', assumes that

> …the mind has a mental concordance of every word it has encountered, a concordance that has been richly glossed for social, physical, discoursal, generic and interpersonal context. This mental concordance is accessible and can be processed in much the same way that a computer concordance is, so that all kinds of patterns, including collocational patterns, are available for use. (Hoey 2005: 11)

Priming, the psycholinguistic phenomenon on which Hoey draws, refers to a well-established effect by which a word (e.g. *nurse*) is recognised faster when preceded by a closely related word (e.g. *hospital*) (for early examples, see Meyer 1971; Scarborough et al. 1977). Such effects have been demonstrated for both paradigmatic relations (e.g. semantically related words, structurally related words) and syntagmatic relations (e.g. words which normally co-occur in a syntactic relationship – *cat* + *sleep*). Not limited to comprehension, studies also show that language users have strong tendency to also reproduce forms which they have recently been exposed to (e.g. Bock 1986; Szmrecsanyi 2005).

In psychology, such priming effects are often accounted for by 'spreading of activation' (e.g. Collins & Loftus 1975). Recognizing a word involves activating relevant mental representation of that word. Perception of the prime word

automatically leads to activation spreading across a network of connections to closely related words, thus reducing the threshold for evidence needed to activate those words. That related words facilitate activation, as evidenced by faster reaction times, suggests that mental representations of words are not discrete, but rather aspects of representations are shared or overlap.

Sentence processing studies demonstrate how language users are able to utilize such cues in language processing. For example, using eye-tracking McDonald and Shillcock (2003a, 2003b) investigated the processing of bigrams differentiated on the basis of *transitional probabilities* derived from the British National Corpus (BNC). This is the simple probability that word *B* will follow word *A* in a bigram *AB* (i.e. Frequency of *AB* / Frequency of *A*) and is analogous to corpus linguistic approaches to collocation. They demonstrate that in stimuli such as *The dictionary editor tried to resolve disputes/meanings in a relatively impartial way*, transitional probabilities (*resolve disputes* = high probability; *resolve meanings* = low probability) have a significant effect on the duration of fixations on the target noun - longer fixations being shown for low probability transitions. They argue that in language comprehension we are able to able bring to bear statistical information about the likelihood of upcoming input based on our prior experience.

For second language learners target-like mastery of collocation norms is especially difficult (Wray 2000). Given differences between the way in which a first language and a second language are acquired, this is hardly surprising. For example, second language acquisition takes place after the linguistic conventions of the first language are already entrenched in the mind; the language data to which learners are exposed differs in terms of both quantity and quality; and, in an instructed context, grammar and lexis are often taught separately.

A sizable body of research documents differences in collocations produced by learner learners and native-speakers (see Nesselhauf 2005 for a review). Much research has focused on collocation 'errors' – i.e. learner collocations which deviate from target language norms. Errors might arise for various of reasons. For example, learners may transfer the lexical or grammatical patterning of words from their native language – e.g. *make* a photo* for *take a photo* in the case of German leaners of English. Or, learners may simply lack knowledge of combinatory restrictions, which are, to a degree, arbitrary – e.g. *plastic operation** for *plastic surgery* (Siyavona & Schmitt 2008).

It is also well documented that second language learners have a tendency to overuse non-restricted words which can combine readily (e.g. Hasselgren 1994) – e.g. *good*, *big*, *bad*, *make*, *take* etc. In the absence of knowledge of a more precise choice, such words provide a useful strategy, and are described by Hasselgren (1994: 250) as 'lexical teddy bears' – "core words […] learnt early, widely usable, and above all safe (because they do not show up as errors)".

How might differences between native speaker and learner use of collocation impact on language processing? Howarth (1998: 162) suggests that learner collocation errors "can have an appreciable impact on the effectiveness of a piece of writing, and their cumulative effect can be a serious loss of precision." Millar (2011) explored how learner collocation errrors impact on processing by native speakers. He demonstrated that 'malformed' learner collocations take longer to read by native speakers and disrupt processing.

While it seems that collocation errors have potential to impede communication, it is unclear how learner reliance on weak collocations that are not errors (so-called lexical teddy bears) might affect processing. Because lexical teddy bears often lack the predictive power of a more precise collocate (e.g. *nice memories* vs. *fond memories*), according to Lexical Priming, the use of such non-restricted choices has implications for how the subsequent words are processed. A reading experiment was undertaken in order to explore how this aspect of learner collocation usage. Specifically, the experiment sought to measure how learner use of a general collocate in place of a more precise option (e.g. *good memories* vs. *fond memories*) impacts on processing by native speaker readers.

METHODS

Hypothesis

On the paradigmatic level, high frequency, general adjectives (e.g. *nice*, *good*, *bad*), which combine to form 'weak' collocations (e.g. *nice memories*), are likely to be processed quickly by the reader – word frequency is one of the most important determinants in lexical decision and word-naming tasks (e.g. Forster & Chambers, 1973). On the other hand, low frequency, more precise adjectives (e.g. *fond*, *vivid*, *cherished*), which can combine to form 'strong' collocations, will take longer to process. While the adjective *nice* has little predictive power over *memories*, we might expect the word *fond* to prime *memories*. Thus, on the syntagmatic level, strong collocates are expected to 'prime' readers and therefore facilitate the processing of the subsequent noun (e.g. *memories*). It is therefore hypothesised that words preceded by a 'strong' (i.e. restricted) collocate are likely to be processed faster than the same words preceded by a 'weak' (i.e. non-restricted) collocate.

Corpus data

All collocations tested in the experiment were naturally occurring samples drawn from language corpora. The one hundred million words of the British National Corpus (BNC) served as the source of native speaker data, and learner data came from the Longman Learner Corpus (LLC) – this contains c. nine million words written by learners of English from a range of proficiency levels. Using part of speech annotation, bigrams tagged as 'adjective + noun' were extracted from the learner corpus. Two types of collocation were identified: (1) 'weak' learner collocations; and (2) semantically similar 'strong' native speaker collocations. The notion of collocation strength was opertationalised through Mutual Information

(MI) scores calculated for adjective-noun bigrams in the BNC. By comparing the observed co-occurrence of two words in the text with the predicted co-occurrence if the all words in the text occurred in random order, MI provides a measure of the amount of information one random variable contains about another (Manning & Schütze 1999: 66), and is often interpreted as the strength of the association between two words in a corpus (McEnery et al. 2006). Firstly, MI scores were calculated for all adjective noun bigrams with a frequency greater than 5. These values were then mapped onto the learner data so that MI of bigrams in the BNC served as the common measurement of collocation strength in both datasets.

'Weak' collocations in the learner dataset were defined as adjective-noun bigrams with an MI score of lower than 3 – this threshold corresponds to the first quartile of MI scores in the learner data set. 'Strong' collocations were defined as adjective-noun bigrams in the BNC with an MI score of higher than 8 – this threshold corresponds to the third quartile in this data set. The two distributions of MI scores are shown in Figure 1 – 'weak' collocations are represented by shaded area (1) and 'strong' collocations by (2).

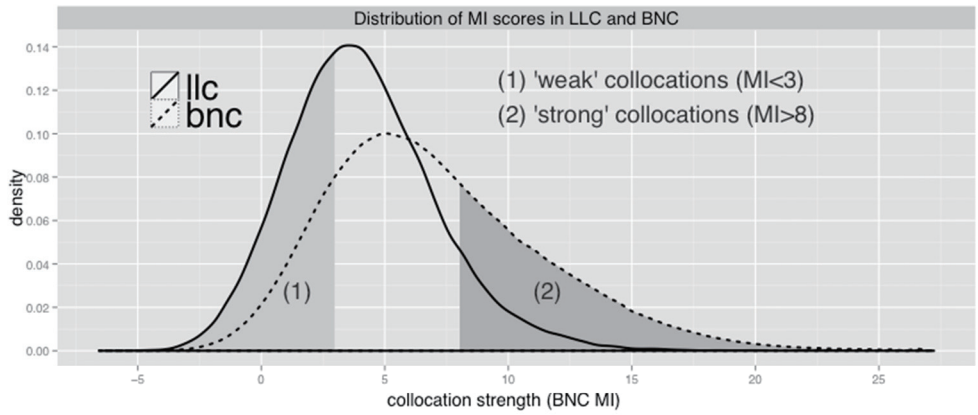


Figure 1: Distribution of MI scores

From the lists of weak and strong collocations, bigram pairs which differed only in terms of the adjective and were deemed to be semantically similar were retained as candidates for use in the self-paced reading experiment – e.g. *bad/adverse effect*, *nice memories/fond memories* and *good/friendly atmosphere*.

Materials

In total, 38 bigram pairs were selected for use in a self-paced reading experiment. The self-paced reading task is based on the premise that eye movement data can reflect moment-to-moment cognitive processes in reading (Rayner 1998). This technique essentially involves holding the gaze still and moving the text into the centre of vision—a task which, compared with eye-tracking, is technically less demanding. Typically, a sentence is presented one word at a time on a computer

screen with the reader controlling the pace by pushing a button. Divergences in mean reading times for each word between conditions provide quantitative evidence for differences in attention deployed in reading particular words (see Rayner 1998). The bigram pairs were embedded in contrived sentences so that each sentence had two conditions: (1) a weak adjective-noun learner collocation; and (2) a semantically similar strong adjective-noun collocation. The conditions, thus, differed only in terms of the adjective – e.g. *Jane had many* ***good/fond memories*** *of childhood holidays spent in Blackpool.* Sentences were 11 words in length, and collocations were embedded as the fourth and fifth word, as the direct object of the main verb, and the context did not overtly predict the bigram. All stimuli used in the experiment are listed in the Table 1.

Table 1: Stimuli used in the Experiment

'Weak' vs. 'strong' collocations – experimental items in **bold**, condition 1 in text, condition 2 in parentheses
1. Changes had a **bad (adverse) effect** on employment and the local economy.
2. Sophie had some **big (ambitious) plans** for the relaunch of the shop.
3. Alex had a **big (heated) argument** with his boss about his wages.
4. Dan watched a **big (heated) debate** on TV about the economic problems.
5. Jim wanted the **big (lucrative) contract** with the government for his company.
6. Tom made a **big (momentous) decision** when he finally quit his job.
7. Lucy had a **big (profound) influence** on her family, students and colleagues.
8. Jim got a **clear (comprehensive) guide** explaining how to make an application.
9. Alice gave a **free (complimentary) bottle** of champagne to all new customers.
10. Megan had a **good (competetive) advantage** over other players in the tournament.
11. Adam offered a **good (convincing) explanation** for the changes that he observed
12. Anna made a **good (determined) effort** to finish the project on time.
13. Alan had a **good (distinguished) career** as a politician and business leader.
14. Jake chose a **good (five-star) hotel** as the venue for the conference.
15. Amy had many **good (fond) memories** of childhood holidays spent in Blackpool.
16. Alice gave a **good (impassioned) speech** about poverty in the developing world.
17. Ron praised the **good (meticulous) planning** that had gone into the event.
18. Jack had many **good (redeeming) features** which made him a popular guy.
19. Paul liked the **good (friendly) atmosphere** and recommended the restaurant to friends.
20. Greg had a **good (sound) basis** for the choices that he made.
21. Kate enjoyed the **good (spectacular) views** from the top of the mountain.
22. Ken had a **good (uncanny) ability** of knowing what people were thinking.
23. Katie gave a **good (valuable) insight** into the way the company functioned.
24. Tony took a **good (well-earned) break** after the project had been finished.
25. Jamie made some **great (sweeping) changes** to the organisation of the company.
26. Residents had a **great (urgent) need** for drinking water after the floods.
27. Tyler described the **hard (stark) reality** of life in a war zone.
28. James had a **large (profound) effect** on almost everyone whom he met.
29. Ian saw some **large (significant) differences** in the way people treated him.
30. Gary had a **nice (well-kept) garden** that many of the neighbours admired.

31. Jane used her **old (maiden) name** when she filled in official forms.
32. The item had **personal (sentimental) value** to Jane and other family members.
33. Photos showed the **slow (painstaking) work** involved in making the traditional clothes.
34. Vicki saw a **small (dim) light** which was flickering in the distance.
35. Tom had a **small (niggling) doubt** about the decision he had made.
36. Bruce had a **small (younger) brother** who was still living at home.
37. Sarah had a **special (flamboyant) style** which made her popular on TV.
38. July was a **special (turning) point** in the government's general election campaign.

Design

A between-groups design was used. Two lists were created, each comprising 38 experimental items, with the conditions counter-balanced so that each list contained either the weak or strong version of each item. Participants were randomly assigned to one of the two groups. Each participant, thus, read all sentences, but saw only one condition for each experimental item. The numbers of participants in each group were balanced. The order of stimuli was randomized across participants to minimize any ordering effects. To encourage a focus on meaning, each sentence was followed by a simple true/false statement relating to its content – e.g. *Jane didn't enjoy her childhood holidays in Blackpool* (false).

Participants

In total 44 students in the second or third year of undergraduate courses in the Department of Linguistics and English Language at Lancaster University in the UK took part in the experiment. All participants were native speakers of British English and were recruited in accordance with institutional guidelines.

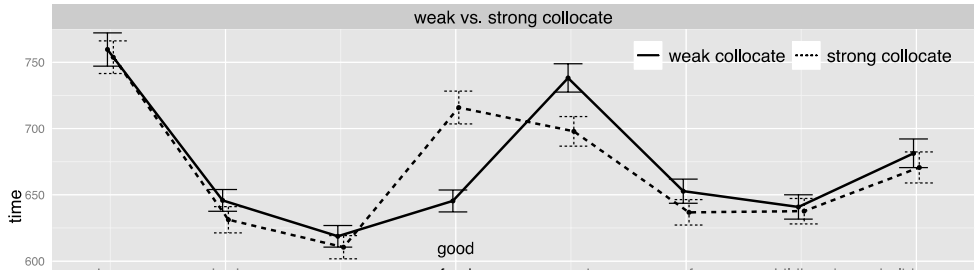
Equipment and procedure

The experiment involved word-by-word self-paced reading with each word centred on the computer display. The experiment was scripted using *PsyScript* and administered on a 2.4GHz Macintosh computer.[1] Participants were asked to read at a pace that would facilitate answering of a comprehension question, and were provided with ten practice items to familiarise them with the procedure. Participants controlled the pace of presentation by clicking on an external mouse. The times between clicks were recorded, as were the participants' answers to true/false questions.

RESULTS

On average each experiment lasted approximately fifteen minutes. Comprehension accuracy over the two experimental conditions was comparable – 91.6% in the weak collocation condition and 93.1% in the strong condition (the difference between

these means is not statistically significant). Figure 2 shows the mean reading times per word from the start of the sentence up until three words after the experimental items. Up until the onset of the manipulation (*A*), reading times between conditions are closely aligned. On average, weakly collocating adjectives were read 71ms faster than strongly collocating adjectives. However, the target nouns (*B*) were read on average 21ms faster when preceded by the strong collocate.



word position	A–1	A	B	B+1	B+2	B+3
weak collocate – mean (s.e.)	605 (8.5)	707 (11.8)	695 (10.7)	637 (9.4)	637 (9.3)	664 (11.3)
strong collocate – mean (s.e.)	606 (7.7)	636 (8.2)	726 (10.4)	648 (8.9)	636 (8.8)	677 (10.8)
effect size (Beta weight)	–	–0.0037***	0.0011*	–	–	–

Figure 2: Mean reading times of weak/strong collocations

In order to account for a positively skewed distribution, reading times were converted to natural logarithms and analysis was carried out using a linear mixed effects model. The explanatory variable condition (weak vs. strong) was submitted as a fixed effect and participants and item as random effects. This model was fitted for each position of the bigram (i.e. *A* and *B*). Standardized regression coefficients (ß, calculated following Bring 1994) are given as a measure of effect size. A statistically significant condition effect was observed on the adjective position, *A*(b=-0.099, *SE*=0.013, t=6.55, p<.001, ß=-0.0037), and on the target noun, *B*(b=0.030, *SE*=0.014, t=2.15, p<.05, ß=0.0011).

A second model was fitted in which the dichotomous explanatory variable condition was replaced by the following continuous variables: word frequency (BNC frequencies logarithmically transformed); and the MI of the bigram (also based on data from the BNC). For the adjective (*A*) word frequency was submitted as a fixed effect – this significantly predicted reading times (b= 0.017, *SE*= 0.0024, t= 7.021 , p<0.05). For the noun (*B*), in addition to word frequency, MI was submitted as an additional fixed effect. The MI of the bigram had a statistically significant effect on the reading time of nouns (b= 0.004, *SE*= 0.002, t= 2.101, p<0.05), while no significant effects were observed for word frequency.

DISCUSSION

The analyses above indicate that the time required to process the adjective in the bigrams was primarily driven by word frequency. On the subsequent noun, where

frequency did not differ across conditions, reading times of nouns was significantly influenced by the strength of the collocation with the preceding adjective. As hypothesised, nouns preceded by a weak collocate tended to take slightly longer to process than when preceded by strong collocate. This finding is consistent with evidence from other sentence processing studies (e.g. McDonald & Shillcock 2003a, b), and is interpreted as support for aspects of Hoey's (2005) theory of Lexical Priming.

Consider the sentence *Jane had many **good memories**…* vs. *Jane had many **fond memories**….* Although both *good memories* and *fond memories* are acceptable combinations, the amount of attention required to process *memories* appears to be influenced by the preceding adjective. While in both conditions the context prior to *memories* primes reader to expect a noun, the strength of primings about which noun are likely to differ. While the adjective *good* can co-occur with many different noun types (in the BNC c. 4,800), it co-occurs with *memories* only 15 times in the BNC. As this corresponds to a probability of close to zero (see discussion of transitional probabilities above), *good* has extremely little predictive power over *memories.* However, the adjective *fond* is heavily restricted in the number of types with which it can co-occur. In the BNC it co-occurs with only 58 noun types, and over one third of these co-occurrences are with a single type: *memories*. We might infer that at some level, the adjective *fond* primes reader to expect the word *memories* (transitional probability = 0.433)*,* and that these 'lexical primings' are reflected in the weak condition effect which we observe on the target noun.

It is therefore suggested that readers draw on expectations about the upcoming input. How realistic is this? Non-linguistic cognition clearly involves predictive processes. For example, by observing the movements of an opponent, an experienced tennis player might anticipate the trajectory of the return shot and adjust his or her position accordingly (Wilson & Knoblich 2005). Compared to a purely reactive system, predictive processing has clear benefits for many cognitive processes – e.g. perception, motor control, decision-making, theory of mind (Bubic et al. 2010). Language comprehension too seems to involve predictions based on both linguistic and non-linguistic cues (see Kamide 2008 for a review). As a behavioural system, language lends itself to predictions: we are experienced as both producers and comprehenders, and the system is predictable on many linguistic levels – e.g. phonotactics, syntax, discourse. It seems likely that collocation observed in text corpora reflects cues on which the comprehender can also draw.

Previous studies indicate that many learners tend to rely heavily on lexical teddy bears combing to make weak collocations. Although this represent a useful communication strategy, the findings of this study indicate that learners' use of weak collocations can have a measurable effect on processing by the addressee. While learner reliance on weak collocations may not disrupt processing in the same way as errors (see Millar 2011), the findings from this study point to a subtle effect on the reader which may in turn impact on communication.

We might equate the effect of collocation strength to the linguistic notion of *cohesion* (specifically, *lexical cohesion*) – the way in which meanings within a text relate with each other. Halliday and Hasan (1976: 284) argue that cohesion is, among other things, achieved by means of collocation: what they describe as "the association of lexical items that regularly co-occur". They argue that cohesive features can lead to coherence – i.e. the degree to which a text makes sense and is logical to a reader. Weak association between lexical items in a text might, therefore, lead to a feeling, either conscious or subconscious, that the writing does not makes sense and/or is not logical. While the effect of single weak collocations is likely to be small, as with collocation errors, the cumulative effect may be more tangible.

In language teaching and SLA research there seems to be a general consensus that instruction should ensure that learners develop a rich repertoire of formulaic sequences, including collocation (Ellis, 2005). Many such claims would appear to be based on the (often implicit) assumption that "the most efficient and successful communication will occur where the speaker and hearer are very familiar with each others' speech patterns, or indeed share the same micro-variety" (Wray, 2002: 479). With regards to writing, rather than speech, the findings presented here seem to support this position.

It is important to stress that any conclusions concerning the impact of learner collocation use on communication need to be contextually restricted. Second language communication need not imply learner interactions with a native speaker – all participants in this study were native speakers of English. This is most evident in situations where English is used as a lingua franca between speakers for whom English is not the first language (cf. Seidlhofer 2001). If and how collocation, and other aspects of formulaicity, support processing in the context of English as a lingua franca is an interesting empirical question and an area for future study.

Finally, it is important to note some limitations of this study. Firstly the instrument used here is a substitute for natural reading. In natural reading about 10–15 per cent of all eye movements are in fact backwards (Rayner and Pollatsek 1994). These regressive saccades, or regressions, are often made when readers have difficulty processing the text (Rayner 1998). However, self-paced reading constrains movement to a left-to-right forward motion, in which participants are not able to make regressions. The generalisability of the findings is also limited by the relatively small number of items tested. All items used in the experiment were attested samples of natural language drawn from corpora. Although this arguably contributed to the authenticity to the stimuli, the total of number of items available was limited (only 38).

CONCLUSION

Integrating corpus data and experimentation, the present study has explored how learner use of so-called Lexical Teddy Bears are processed in reading by native

speakers. The results indicate that when native speakers read these non-predictive choices they take marginally longer to process the next word. These findings are in line with recent usage-based theories of language, and provide additional empirical support for one of the core claims of Lexical Priming theory (Hoey 2005: 13) – specifically, that language users are "primed to expect words to occur with certain other words (their collocates)". It also has been suggested that collocation has potential to impact of communication in that it can contribute to cohesion and predictive cues on which the comprehender can also draw. This study, thus, adds to empirical psycholinguistic evidence that collocation (and by extension formulaic sequences in general) can, in certain contexts, contribute to a learner's communicative competence.

Notes

[1] Published by the Psychology Department, Lancaster University (http://www.psych.lancs.ac.uk/software/)

REFERENCES

Bock, K. 1986. 'Syntactic Persistence in language production,' *Cognitive Psychology* 18: 355-387.

Bring, J. (1994). 'How to Standardize Regression Coefficients,' *The American Statistician* 48/3: 209-213.

Bubic, A., D. Y. von Cramon, and **R. I. Schubotz.** 2010. 'Prediction, cognition and the brain,' *Frontiers in Human Neuroscience* 4:1–15.

Collins, A. M., and **E. F. Loftus.** 1975. 'A spreading-activation theory of semantic processing,' *Psychological Review* 82: 407–428.

Ellis, R. (2005). Principles of instructed language learning. *System*, 33/2, 209-224.

Foster, K. and **S. M. Chamber.** 1973 'Lexical Access and Naming Time,' *Journal of Verbal Learning and Verbal Behavior* 12: 627-635.

Halliday, M. A. K., and **R. Hasan.** 1976. *Cohesion in English*. London: Longman.

Hasselgren, A. 1994. 'Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary,' *International Journal of Applied Linguistics* 4/2: 237-258.

Hoey, M. 2005. *Lexical priming: A new theory of words and language*. Routledge.

Howarth, P. 1998. 'The phraseology of learners' academic writing' in A. P. Cowie (ed): *Phraseology: Theory, analysis, and applications*. Clarendon Press, pp. 161-168.

Kamide, Y. 2008. 'Anticipatory processes in sentence processing,' *Language and Linguistics Compass* 2/4: 647-670.

Manning, C. D., and **H. Schütze.** 1999. *Foundations of statistical natural language processing*. MIT Press.

McDonald, S. A., and **R. C. Shillcock.** 2003a. 'Eye movements reveal the on-line computation of lexical probabilities during reading,' *Psychological Science* 14/6: 648-652.

McDonald, S. A., and **R. C. Shillcock.** 2003b. 'Low-level predictive inference in reading: the influence of transitional probabilities on eye movements,' *Vision Research* 43/16: 1735-1751.

McEnery, T., Z. Xiao, and **Y. Tono.** 2006. *Corpus-based language studies: An advanced resource book*. Routledge.

Meyer, D. E. 1971. 'Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations,' *Journal Of Experimental Psychology* 90/2: 227-234.

Millar, N. 2011. 'Processing malformed formulaic language,' *Applied Linguistics* 32/2: 129–148.

Nesselhauf, N. 2005. *Collocations in a learner corpus*. John Benjamins.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), Language and Communication. Longman.

Rayner, K. 1998. 'Eye movements in reading and information processing: 20 years of research,' *Psychological Bulletin* 124/3: 372–422.

Rayner, K., & Pollatsek, A. (1994). *The psychology of reading*. Erlbaum.

Scarborough, D. L., C. Cortese, and **H. S. Scarborough.** 1977. 'Frequency and repetition effects in lexical memory,' *Journal of Experimental Psychology: Human Perception and Performance* 3/1: 1-17.

Seidlhofer, B. 2001. 'Closing a conceptual gap: The case for a description of English as a lingua franca,' *International Journal of Applied Linguistics* 11: 133-158.

Siyanova, A. and **Schmitt, N.** 2008. 'L2 learner production and processing of collocation: A multi-study perspective'. *Canadian Modern Language Review*, 64/3: 429-458.

Szmrecsanyi, B. 2005. 'Language users as creatures of habit: A corpus-linguistic analysis of persistence in spoken English,' *Corpus Linguistics and Linguistic Theory* 1/1: 113-150.

Wilson, M., and **G. Knoblich.** 2005. 'The case for motor involvement in perceiving conspecifics,' *Psychological Bulletin* 131/3: 460-473.

Wray, A. 2000. 'Formulaic sequences in second language teaching: Principle and practice,' *Applied Linguistics* 21/4: 463-489.

Wray, A. 2002. *Formulaic Language and the Lexicon*. Cambridge University Press.